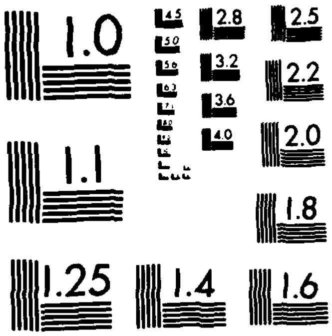
AD-A154 199  REGRESSION DIAGNOSTICS IN PRACTICE -- EXPERIENCES FROM MODELLING JET ENGINE COSTS(U) RAND CORP SANTA MONICA CA J B GARFINKLE ET AL. JUN 83 RAND/P-6896 SBI-AD-F630 665  1/1

UNCLASSIFIED  F/G 14/1  NL

END

FILMED

DTIC




MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-F530665 (2)

AD-A154 199

REGRESSION DIAGNOSTICS IN PRACTICE --
EXPERIENCES FROM MODELLING JET ENGINE COSTS

Jeffrey B. Garfinkle and John L. Birkler

June 1983







P-6896

# REGRESSION DIAGNOSTICS IN PRACTICE --
# EXPERIENCES FROM MODELLING JET ENGINE COSTS

Jeffrey B. Garfinkle and John L. Birkler

The Rand Corporation
1700 Main Street
Santa Monica, CA 90406

Prepared for the 1983 Joint Statistical Meetings, Toronto, Canada

## Preface

This paper describes how regression diagnostics were used to help develop revised cost-estimating relationships for jet engines. The goal was to derive meaningful, yet easy to use models based on an updated collection of few observations and many variables. First, specific criteria were established for selecting explanatory variables. A variety of numerical and graphical techniques were then used to critique candidate models by examining residuals and evaluating the influence of individual engines. The final models are not only intuitively satisfying, but generally provide better predictions and are easier to use than earlier models. Additionally, the user is provided with a greater understanding of the design and sensitivity of the models, and therefore a better understanding of the actual estimates.

This research was undertaken as part of the "Air Force Resource Financial Management Issues for the 1980s" project of Rand's Project AIR FORCE.

## Introduction

The propulsion system of most new military aircraft is a turbine engine. These complex engines can cost over a billion dollars for development and product improvement. In fact, the propulsion system accounts for as much as 25% of the flyaway cost of a fighter aircraft. Consequently Rand has, over the years, developed mathematical models for use in the early planning stages of engines, which provided estimates of development and production cost. One of the most popular sets of these cost-estimating relationships (CERs) have been incorporated into a Rand developed computer model of aircraft system costs known as DAPCA (Development and Procurement Cost of Aircraft) [1].

The DAPCA model estimates engine costs using the results from two regression equations. Experience has shown the estimates to be quite sensitive to small changes in the data used to drive the equations, and likely to underestimate costs for the latest high performance engines. A revised set of CERs has been developed by Birkler, Garfinkle, and Marks [2] which not only incorporates data from engines developed since the DAPCA equations were derived, but also takes advantage of recently available statistical and computational tools.

This paper focuses on one of these new engine equations as a way of describing our experiences with these tools.

## A Typical CER

The Model Qualifications Test (MQT) is a series of tests used to demonstrate an engine's suitability for production. It is a major milestone in the development of aircraft turbine engines, marking the

transition between development and production. For conceptual planning studies, preliminary tradeoff analyses, etc., it is desirable to have an easy to use procedure for estimating within $\pm$ 25% the cost of a proposed engine through the successful completion of MQT. An equation that produces this estimate should require few input parameters, all of which should be available during the concept formulation stage -- before blueprints, orders for material, and other engine components can be specified. Also, the predictive capabilities of the model should be well understood. Finally, the signs of the coefficients should be consistent with intuitive notions of the relationship between technology and cost. A CER that has these features will not only have a variety of applications but will better stand the test of time.

## Candidate Explanatory Variables

Our search for suitable explanatory variables began with the hypothesis that the cost of an engine is a function of (1) the size of the engine, (2) the level of technology/performance incorporated into the engine, and (3) the time during which the engine is produced [3]. We also required that a candidate variable be logically related to cost and be available (with a reasonable amount of accuracy) early in the planning cycle. Our final set of candidate explanatory variables that satisfied these criteria is shown in Table 1. Our dependent variable is the cost of development up to successful completion of MQT (MQTDEVCOST).

Table 1

CANDIDATE EXPLANATORY VARIABLES

| Size | Performance/Technology | Time |
|---|---|---|
| Thrust* | Turbine inlet temperature* | Time of Arrival* |
| Weight | Thrust to weight ratio* | |
| Airflow | Mach number* | |
| | Total pressure | |
| | Specific fuel consumption | |
| | Thrust per pound of airflow | |

*These variables may be easier to obtain in a long range planning study than others.

## Cost Data

Data on sixteen turbojet and turbofan engines were collected from those companies who developed and produced the various aircraft turbine engines. The data were adjusted so that comparisons could be made among engines based on constant dollars, like quantities, and for generally similar developmental strategies.

## Transformations of the Data

The specific form of a CER (whether it is linear or logarithmic in the independent and dependent variables) implies various assumptions about technology and cost trends [4]. For example, a model that is linear in both the independent and dependent variables (linear-linear) implies a constant relationship between cost and technology, whereas a linear-log model implies an acceleration in cost. Although it may seem unreasonable to expect either relationship to hold indefinitely over time, these forms could describe certain phases of engine technology. A linear-log model implies a deceleration or an acceleration in cost depending on the values of the coefficients of the logarithmic terms. The log-log model implies a deceleration in cost, as might be expected in a mature technology. To help answer the question of which equation form best describes the actual relationship between engine cost and engine characteristics, all four forms were studied using a variety of diagnostic tools.

## Diagnostic Tools

Our first step in developing a CER for MQTDEVCOST was to compute all possible linear least-squares regressions [5] using from one to

five of the candidate independent variables. Advances in computer algorithms helped to hold this procedure down to a reasonable cost [6]. Mallow's C(p) [7], the multiple correlation coefficient, and our knowledge of propulsion engineering, were all used in selecting a small subset of these models for further analysis.

The next step was to compute full sets of statistics for the candidate models. We also graphically compared the residuals from each model to its fitted values and to its independent variables; partial regression plots were used to provide further indications of the behavior of individual data points. The candidate models were also examined for evidence of collinearity by reviewing the condition index of the data matrix and the decomposition of the estimated regression coefficient variances [8].

Furthermore, a variety of influence measures were used in our search for the "best" model, especially those described by Belsley, Kuh, and Welsch [8] and Cook [9]. These measures indicate the effects of deleting an observation from the data base. In general, we considered an engine influential if its deletion resulted in large changes in various characteristics of the model.

## The Preferred Equation

The preferred equation for estimating engine development costs through MQT is displayed both numerically and graphically in Figure 1. The relationship shown there has three explanatory variables, all of which have intuitive appeal as well as statistical significance [10]. The points on the plot represent the sixteen engines in the data base. Costs for engines above the 45-degree line have been overestimated

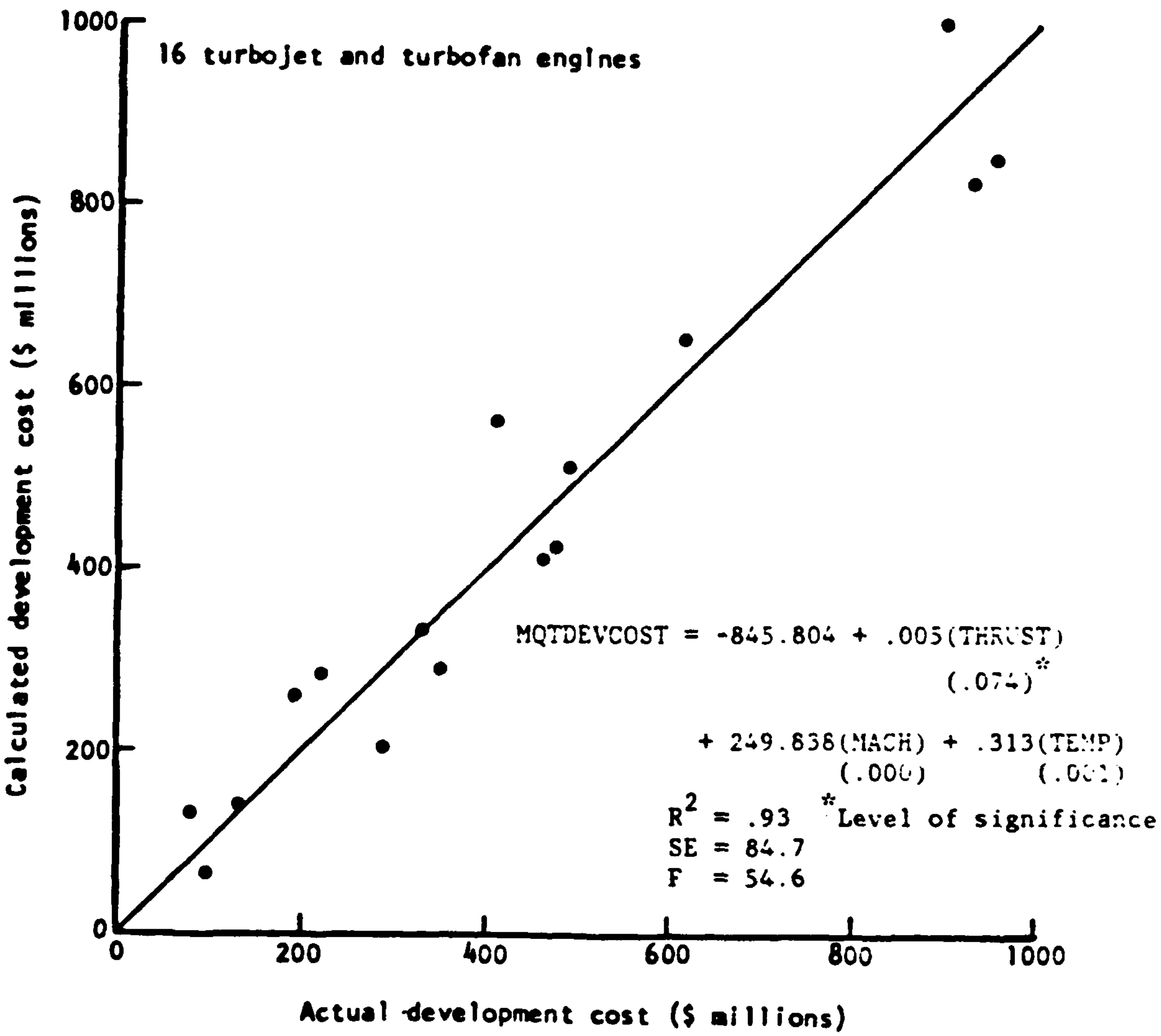



Figure 1

DEVELOPMENT COST (MQT), PREFERRED EQUATION

while costs below the line have been underestimated. A similar plot was used with other models in order to detect outliers and to examine possible trends in the distribution of the residuals.

The studentized residual, hat diagonal, covariance ratio, DFFITS, DFBETAS, and Cook's distance for each data point used in the model are shown in Table 2. Before beginning the anaylsis we expected the F100, the TF39, and the J58 engines to be flagged by these statistics. The F100 should be an influential observation because it is the most technically advanced engine in the sample, having the highest turbine inlet temperature and thrust-to-weight ratio. The TF39 should also be identified as an influential observation because, although it is a subsonic engine, it has the highest thrust rating of all the engines in the sample. The large thrust output is due to its large size and the fact that much of its thrust is generated in a mode quite different from the other engines in the sample. Also, it is the only large, transport type of engine in the data base. The third expected outlier is the J58. This engine is the only engine in the sample designed for a high altitude, high speed reconnaissance mission, which requires a considerably different design and testing approach.

A routine analysis of the residuals only identified the TF30 as being an outlier. The influence measures did a better job in this regard. For many of the diagnostics the F100 was flagged as being an important data point in influencing the regression coefficients. In addition, several of the diagnostics identified the TF39 as being potentially different from other engines in the sample. The J58 had the highest Cook's distance and conspicuously exceeded the extreme values for DFFITS and two of the four DFBETAS.

Table 2

REGRESSION DIAGNOSTICS FOR MQTDEVCOST EQUATION

| Engine | Residual | Hat Diagonal | Covariance Ratio | DFFITS | DFBETAS Intercept | DFBETAS THRMAX | DFBETAS MACH | DFBETAS TEMP | Cook's D |
|---|---|---|---|---|---|---|---|---|---|
| F100 | 1.6098 | 0.2799 | 0.8439 | 1.0037 | -0.7694 | -0.3440 | 0.3494 | 0.6645 | 0.222 |
| F101 | 1.5470 | 0.2441 | 0.8526 | 0.8791 | -0.5387 | 0.0824 | 0.0038 | 0.4850 | 0.173 |
| F404 | -0.5172 | 0.2857 | 1.8011 | -0.3271 | 0.2544 | 0.2036 | -0.0135 | -0.2795 | 0.028 |
| TF30 | -2.1377 | 0.0998 | 0.3920 | -0.7120 | -0.0225 | 0.0019 | -0.4270 | 0.1101 | 0.098 |
| TF33 | -0.1444 | 0.2300 | 1.8255 | -0.0789 | -0.0593 | -0.0415 | 0.0454 | 0.0392 | 0.002 |
| TF34 | -0.9518 | 0.3640 | 1.6206 | -0.7216 | 0.3031 | 0.3922 | 0.4100 | -0.5413 | 0.131 |
| TF39 | -0.4426 | 0.7001 | 4.4006 | -0.6762 | -0.0924 | -0.4758 | 0.4509 | 0.0095 | 0.123 |
| J52 | 0.7383 | 0.1315 | 1.3470 | 0.2898 | 0.1369 | -0.0687 | 0.1069 | -0.1147 | 0.022 |
| J57 | 1.1477 | 0.1146 | 1.0175 | 0.4128 | 0.2705 | -0.0072 | -0.0669 | -0.1670 | 0.042 |
| J58 | -1.6939 | 0.4651 | 1.0477 | -1.5796 | 0.1563 | -0.5458 | -1.1382 | 0.3574 | 0.540 |
| J60 | 0.3775 | 0.2061 | 1.6945 | 0.1923 | 0.0561 | -0.0843 | -0.0828 | 0.0167 | 0.010 |
| J65 | -0.5646 | 0.1430 | 1.4741 | -0.2306 | -0.1300 | 0.0383 | 0.0755 | 0.0548 | 0.014 |
| J71 | -0.8172 | 0.0956 | 1.2371 | -0.2656 | -0.1093 | 0.0713 | 0.0221 | 0.0414 | 0.018 |
| J75 | 0.7089 | 0.3004 | 1.6932 | 0.4645 | 0.3432 | 0.3226 | 0.1176 | -0.3906 | 0.056 |
| J79 | 0.6059 | 0.1225 | 1.4155 | 0.2264 | 0.1196 | 0.0333 | 0.1094 | -0.1311 | 0.014 |
| J85 | -0.0043 | 0.2155 | 1.8054 | -0.0023 | -0.0003 | 0.0013 | -0.0012 | 0.0002 | 0.000 |
| CUTOFF VALUE | ±2.0 | .5 | (.3,1.8) | ±1.0 | ±.5 | ±.5 | ±.5 | ±.5 | .9 |

Note: Statistics that approach or exceed their cutoff values are underlined.

As we initially expected, the F100, TF39, and J58 proved consistently to be the most influential engines in this model. Because future engines are expected to be more like these than the others in the data set, this is an acceptable model.

A large number of other models for estimating engine cost were analyzed before the final set of preferred equations was chosen. The competing models each had at least one of the following drawbacks:

- candidate variables were not significant
- coefficients had counterintuitive signs
- the model had a large estimating error
- independent variables exhibited collinearity
- influential engines were unlike expected future engines.

## Summary

The estimating relationships derived in this project provide improvements in engine cost estimating capablility over the DAPCA equations. Major strong points of these new relationships are intuitive appeal, ease of use, fewer independent variables, and low estimating error. Also, we have insight into the influence of each engine in the data base on the final models. Additional engine development, since the time the DAPCA equations were derived, have yielded useful data that have been added to the data base so that it represents a wider range of engine characteristics.

The results described in this study are intended for estimating the cost of large, modern aircraft engines in the context of long-range planning studies. Any new engine to be estimated must be consistent with the basic assumptions and limited data on which the CERs were

derived. Specifically, the CERs apply to development and pricing practices similar to those of the 1960s and 1970s. The models also require the reasonable assumption that basic gas turbine design in the future will be similar to that of today.

Throughout this study regression diagnostics were used as tools to supplement our understanding of the data. The diagnostics not only identified influential data points that would have otherwise gone undetected, but guided us to a model in which we can believe.

## Footnotes

[1] J. R. Nelson and F. S. Timson, Relating Technology to Acquisition Costs: Aircraft Turbine Engines, R-1288-PR, The Rand Corporation, March 1974. The estimating relationships developed in this research are incorporated in the DAPCA model: H. E. Boren, Jr., A Computer Model for Estimating Development and Procurement Costs of Aircraft (DAPCA III), R-1854-PR, The Rand Corporation, March 1976.

[2] J. L. Birkler, J. B. Garfinkle, and K. E. Marks, Development and Production Cost Estimating Relationships for Aircraft Turbine Engines, N-1882-AF, The Rand Corporation, October 1982.

[3] This hypothesis is more fully discussed in Birkler (1982).

[4] W. L. Stanley and M. Miller, Measuring Technological Change in Jet Fighter Aircraft, The Rand Corporation, R-2249-AF, September 1979, pp. 16-18.

[5] N. Draper and H. Smith, Applied Regression Analysis, 2nd ed., John Wiley and Sons, New York, 1981.

[6] SAS Institute Inc., SAS User's Guide: Statistics, 1982 Edition. Cary, NC:SAS Institute Inc. 1982.

[7] C. Mallow, "Some Comments on C(p)", Technometrics, 15 (1973), pp. 661-675.

[8] D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980.

[9] R. Cook, "Detection of Influential Observations in Linear Regression," Technometrics, 19 (1977), pp. 15-18.

[10] The implications of this set of explanatory variables is more fully discussed in Birkler (1982).

END

FILMED

6-85

DTIC